CSE499A- SENIOR DESIGN REPORT

Comparative Analysis of Pretrained Models for Biomedical Extractive Question Answering Task

Ahmed Ajmine Nehal
2013090 042
ahmed.nehal@northsouth.edu

Farhana Elias Richie
2013820642
farhana.elias@northsouth.edu

Md. Mehedi Hasan
2022107642
mehedi.hasan36@northsouth.edu

Department of Electric & Computer Engineering
North South University, Dhaka

Abstract

A significant advancement in medical science has been made possible by question-answering (QA) systems. The quantity of QA agents working in the medical field has steadily increased recently. In biomedical question-answering applications, answering factoid questions is a crucial task. Its dependability has garnered a lot of interest. Better word representation is crucial for question-answering systems, and word embedding done correctly can boost the system's efficiency considerably. Pretrained models have been widely used in biomedical domains due to their success in general natural language processing tasks. Numerous methods based on pre-trained models have been demonstrated to be successful in biomedical question-answering tasks. We aim to find the best model or models to perform extractive question answering on the bioASQ dataset in the biomedical domain. We will compare different pre-trained models trained on pubMED and PMC corpora. We will fine-tune these for the extractive QA task. With almost the same architecture across the QA tasks, BioMedBERT and BioBERT largely outperforms other state-of-the-art models.

1. Introduction

A growing number of researchers in the fields of natural language processing (NLP) and information retrieval are interested in studying Question Answering (QA) systems. Virtual conversational agents with domain-specific expertise are known as QA agents. Since users occasionally need exact answers to questions in a hurry, they typically prefer QA agents to document readers. Natural language processing techniques, such as calculating term frequency and using it to predict the target word in response to the input query based on historical data, can be used to implement a basic quality assurance agent.

Our research addresses a critical research gap centered around the challenge of achieving optimal accuracy in question-answering (QA) tasks within the biomedical domain. Notably, our focus is on understanding the factors contributing to the observed lower accuracy levels in existing QA models and devising strategies to enhance their performance. Through this investigation, we aim to bridge the existing research gap by proposing innovative approaches and methodologies that have the potential to substantially improve the accuracy of QA tasks in the context of biomedical information. This research is poised to contribute valuable insights, thereby fostering advancements in the development of more accurate and efficient QA models tailored specifically for biomedical applications.

Here, we use the BioASQ dataset. The BioASQ dataset has had a significant impact on the field of biomedical text mining and information retrieval. BioASQ can handle a wide range of biomedical literature. BERT is both powerful empirically and conceptually. Outperforming many task-specific architectures, BERT is the first fine-tuning-based representation model that achieves state-of-the-art performance on a wide range of sentence-level and token-level tasks. Without requiring significant task-specific architecture changes, the pre-trained BERT model can be improved with just one extra output layer to produce state-of-the-art models for a variety of tasks, including question answering and language inference. We used five models: Bert, Biobert, BioMedBert, ClinicalBert, and Bio+ClinicalBert, fine-tuned and compared them.

2. Literature Review

In recent years, advancements in natural language processing (NLP) have pushed the field of extractive question answering (QA) into the cutting edge of biomedical research. Consequently, numerous researchers have conducted experiments aimed at enhancing the ability of NLP models to accurately extract relevant information from biomedical literature or databases in response to user queries.

Chakraborty, S., Bisong, E., Bhatt, S., Wagner, T., Elliott, R., & Mosconi, F. conducted a study introducing BioMedBERT, a BERT variant specifically trained on a massive biomedical literature dataset (BREATHE). It was Pre-trained on BREATHE (200M+ abstracts from PubMed, PMC, etc.) and fine-tuned on BioASQ 6-8 (QA tasks from biomedical literature). The authors demonstrate state-of-the-art performance on EQA tasks in BioASQ benchmarks, surpassing the general-purpose BERT models. Achieved state-of-the-art F1 scores on BioASQ 6-8 EQA tasks (outperforming general-purpose BERT). Improved mean reciprocal rank (MRR) for information retrieval compared to generic BERT models. BioMedBERT's success highlights the importance of domain-specific pre-training for biomedical NLP tasks. Additionally, the authors showcase BioMedBERT's effectiveness in information retrieval, suggesting its broader applicability[9].

In another study, *Xu, G., Rong, W., Wang, Y., Ouyang, Y., & Xiong, Z.* propose a framework to enrich text representations for extractive QA using external features like part-of-speech tags and general named entity recognition. These features complement BERT's contextual understanding, particularly for biomedical terms and syntactic aspects. The model achieves promising results on BioASQ tasks, showcasing the potential of hybrid approaches combining pre-trained models with domain-specific linguistic features. It was Pre-trained on PubMed abstracts and fine-tuned on BioASQ 8 task B (identification of relevant passages). Achieved significant F1 score improvements over the baseline BERT model on BioASQ 8 task B, demonstrating the effectiveness of enriching representations with external features. [7]

Yoon, W., Jackson, R., Kang, J., & Lagerberg, A. explore leveraging sequence tagging models for extractive QA in the biomedical domain. Their proposed model, BioTag-QA, utilizes conditional random fields (CRFs) to identify answer spans within passages, followed by a BERT-based ranking step to choose the most relevant segments. BioTag-QA demonstrates competitive performance on BioASQ, highlighting the efficacy of combining CRF's structured labeling with BERT's powerful representation learning. It was Pre-trained on PubMed abstracts and fine-tuned on BioASQ 7 task A (extractive summarization of factual questions). Obtained competitive F1 scores and ROUGE scores on BioASQ 7 task A, highlighting the efficacy of combining CRF tagging with BERT for answer span identification. [8]

Conducting research gap analysis of three distinct studies conducted by the authors noteworthy insights and potential areas for enhancement in the field of biomedical natural language processing (NLP) have been identified. The author highlights the promise of BioMedBERT, yet suggests that fine-tuning for specific subfields, such as genetics or oncology, could substantially enhance performance in those domains. Additionally, the integration of data from electronic health records and clinical notes is proposed to augment BioMedBERT's understanding of real-world clinical settings and applications in patient

care. The exploration of external features underscores the need for further research to identify optimal feature combinations and develop dynamic feature selection methods tailored to different question types. While the combination of external features with pre-trained language models like BioMedBERT holds potential for improved performance, the study limits their use to independent application, calling for exploration of hybrid approaches. Lastly, the author's focus on single-span answers in biomedical question-answering is acknowledged, with a critical recommendation to explore techniques for extracting multiple spans from passages. The model's reliance on sequence tagging alone is highlighted, suggesting that more advanced reasoning mechanisms, such as attention or dependency parsing, could enhance accuracy by capturing complex relationships between entities and concepts.

In conclusion, these studies collectively underscore the need for specialized fine-tuning, exploration of feature combinations, and the development of advanced reasoning mechanisms to advance the capabilities of biomedical NLP models for extractive question answering.

3. Methodology

3.1 Dataset

3.1.1 Dataset Description

The dataset used in our study is derived from the BioASQ factoid datasets, convertible to the SQuAD dataset format. We have used the 4b dataset for our research, which contains 1307 rows of data. This dataset incorporates complete abstracts identified by PubMed IDs (PMIDs), complemented by associated questions and answers provided by the BioASQ organizers. The BioASQ question answering (QA) benchmark dataset is uniquely designed to mirror the information needs of biomedical experts, presenting a more realistic and challenging scenario compared to many existing datasets. Unlike previous QA benchmarks with solely exact answers. The dataset is dynamic, continuously expanding as the BioASQ challenge progresses, generating new data for ongoing researchers[3][4].

3.1.2 **Dataset Pre-processing**

In the preprocessing of the "BioASQ" dataset for QA tasks, particularly when converting it into the SQuAD format, a series of steps are implemented. The transformation involves structuring the dataset to match the SQuAD format, and establishing connections between

questions and context paragraphs. Strategies such as data augmentation, rephrasing, or selectively removing biased instances are applied during preprocessing to enhance the model's generalization and minimize potential biases. Additionally, standard text preprocessing techniques, including tokenization, stemming, and stop-word removal, are used to ensure compatibility with downstream QA models. **We tokenized the dataset by each model's own tokenizer.** The refined dataset is then ready for training and evaluating QA models, promoting a more standardized and unbiased approach to question answering.

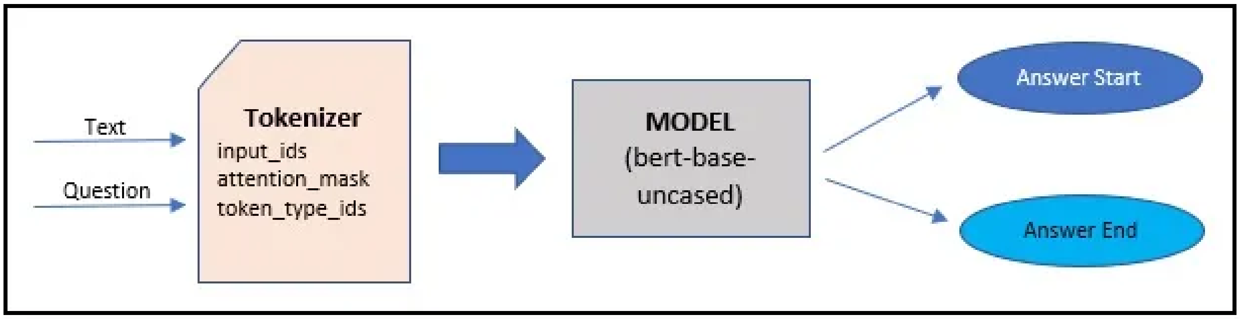


Figure: Dataset Tokenizing

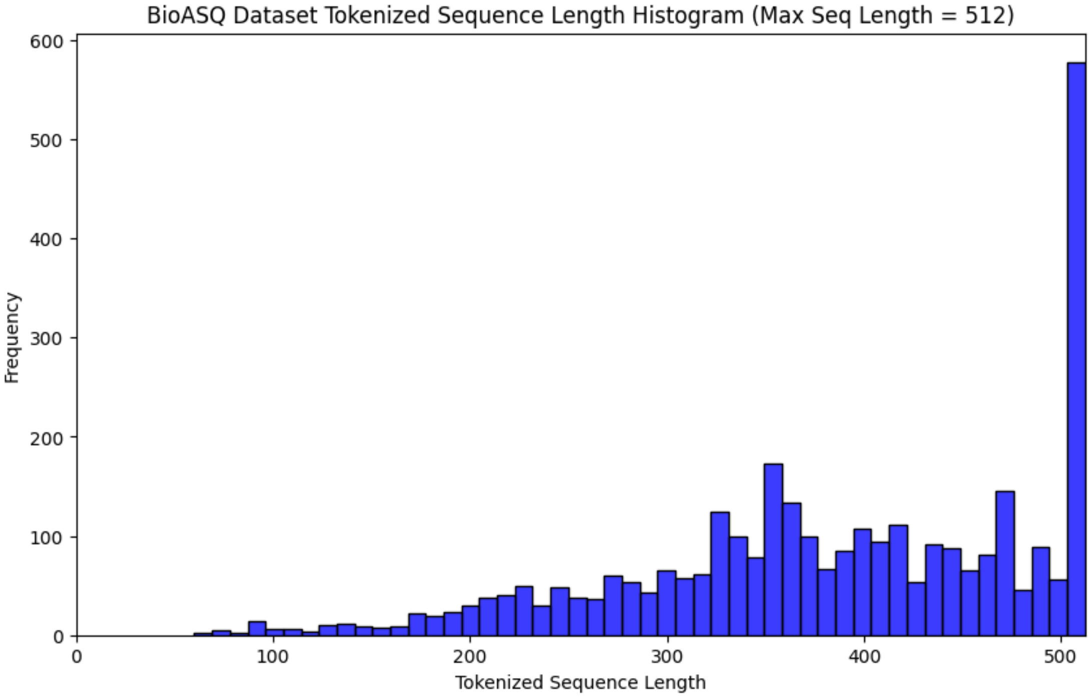


Figure: BioASQ-4b Dataset Tokenized Sequence Length Graph

3.1.3 Dataset Splitting

Dataset splitting involves dividing data into training and testing sets. During training, the model learns patterns and relationships within the data, while the testing set remains unseen until evaluation. **The widely-used train test split function "train_test_split" from the "datasets" library of hugging face is employed for data splitting, allowing us to specify the desired ratio of 80 percent for training and 20 percent for testing.** The split maintains the original data distribution to ensure unbiased evaluation.

Table: Statistics of biomedical question answering datasets

Dataset	Number of questions (Train Set)	Number of questions (Validation Set)	Number of questions (Test Set)
BioASQ 4b-factoid	1046	261	500

3.2 Pre-trained Models

In our study, We conducted a comparative analysis of the performance of five language representation models in the context of fine-tuning for an extractive Question-Answering (QA) task, all designed with a uniform architecture. These models encompass BERT, BioBERT, bio + clinical BERT, ClinicalBERT, BioMedBERT. The consistent architecture facilitates a comprehensive performance comparison across these models. Notably, the inclusion of the BERT model serves as a benchmark for evaluating and contrasting the performance of the other specialized models in our analysis. This systematic approach allows for a rigorous assessment of the relative effectiveness and capabilities of each transformer model within the scope of our study. Brief descriptions of the models are provided as follows:

3.2.1 BERT

BERT is designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both the left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications[1].

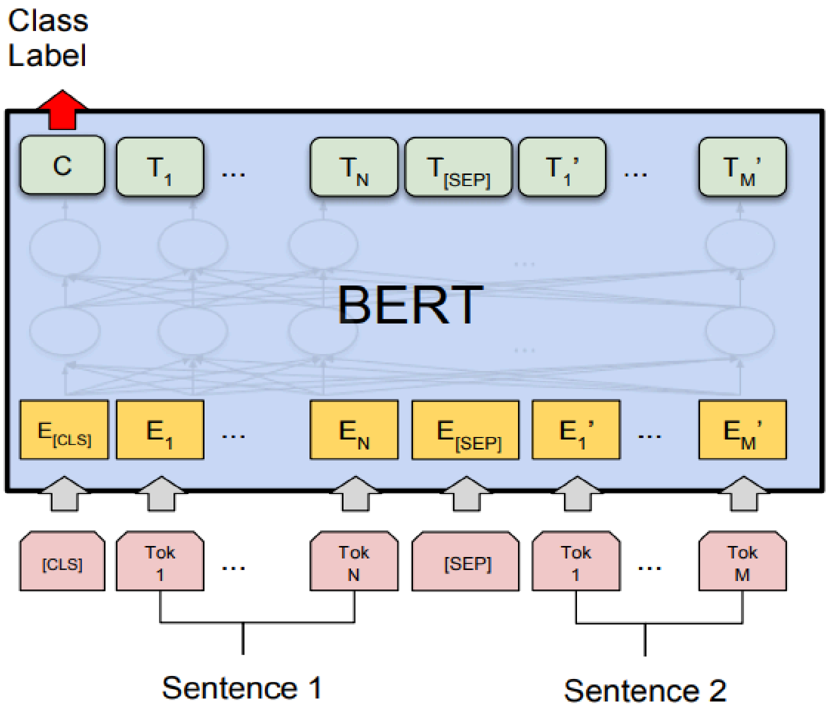


Figure: Bert Models Architecture

3.2.2 BioBERT

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) , is a special language model trained on big medical collections of text. It learned a lot from PubMed and PMC journals. BioBERT has almost the same architecture for different tasks, and it works better than BERT and other top models in various tasks related to understanding and extracting information from biomedical texts[2].

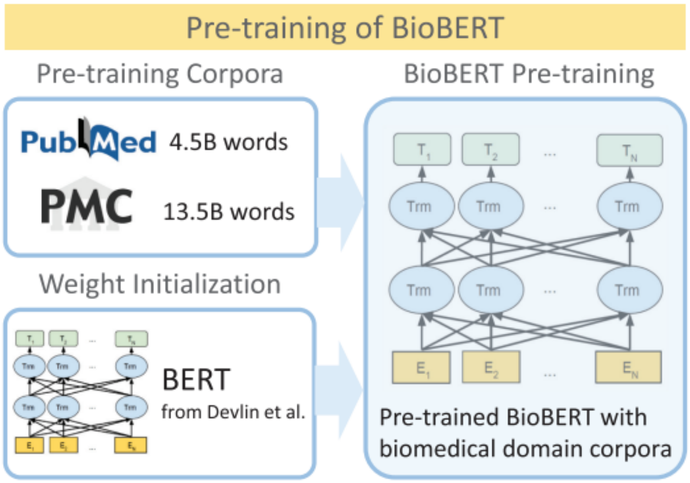


Figure: BioBERT Model Architecture

3.2.3 BioMedBERT

BioMedBERT model developed by Microsoft. This model underwent meticulous pretraining from the ground up, utilizing a diverse dataset comprising abstracts from PubMed and full-text articles from PubMedCentral. This comprehensive training approach aimed to establish a robust understanding of biomedical language nuances. BioMedBERT has demonstrated state-of-the-art performance across various biomedical natural language processing (NLP) tasks, including extractive question answering, highlighting its significance in the field[10].

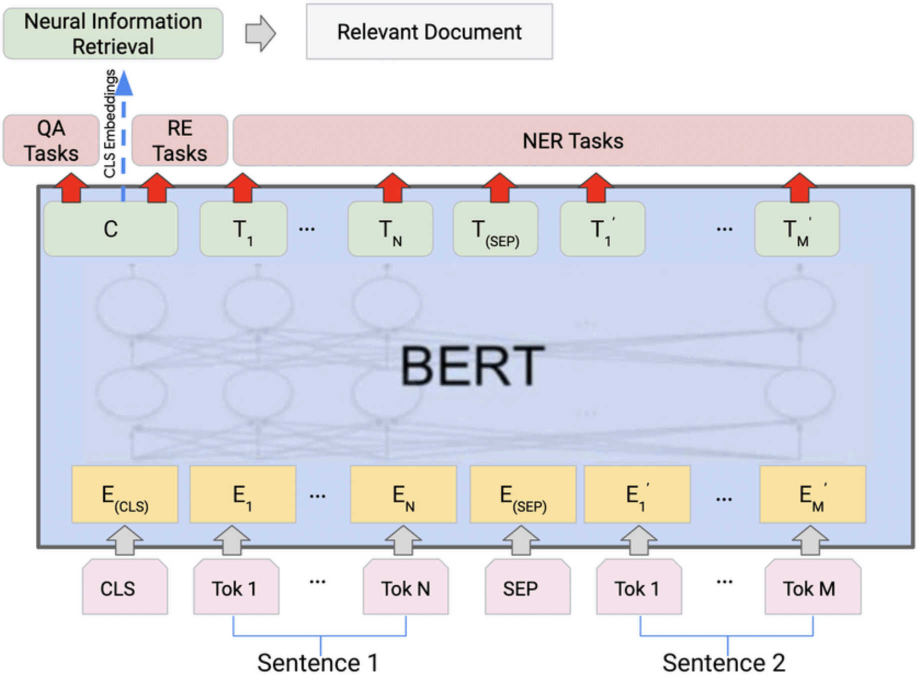


Figure: BioMedBERT Information Retrieval Architecture

3.2.4 ClinicalBERT

ClinicalBERT is a pre-trained language representation model specifically designed for clinical text in the medical domain. It is an extension of the BERT. ClinicalBERT is trained on large-scale clinical text data to capture domain-specific semantics and contextual information present in electronic health records (EHRs), medical literature, and other clinical documents. This fine-tuning of medical data allows ClinicalBERT to better understand the nuances and language intricacies specific to the healthcare domain[6].

3.2.5 Bio+ClinicalBERT

Bio+Clinical BERT is a powerful language model specifically designed for the healthcare domain. It builds upon the foundation of two existing models: BioBERT and regular BERT. Bio+Clinical BERT combines the strengths of both models. It starts with the pre-trained weights of BioBERT and then further refines them by training on a vast collection of electronic health records from MIMIC III, a database of ICU patient notes[5].

Table: Corpora of Pre-trained Models

Pre-Trained Models	Corpora
BERT	BookCorpus: A broad base of general domain knowledge. English Wikipedia: Factual information and relationships between entities in diverse topics.
BioBERT	PubMed Abstracts, PubMed Central Full Texts: Biomedical and life sciences articles. Species-800, BC2GM corpus: NER for species mentions in scientific texts, and biomedical entities.
BioMedBERT	BREATHE 1.0: The variety ensures broad coverage of biomedical topics and knowledge types. PubMed Abstracts, PubMed Central Full Texts: Biomedical and life sciences articles provide extensive factual information and research insights. Species-800, BC2GM corpus: NER for species mentions in scientific texts, and biomedical entities.
ClinicalBERT	MIMIC-III: The massive dataset of electronic health records from ICU patients at Beth Israel Hospital provides domain-specific clinical context.
Bio+ClinicalBERT	MIMIC-III(MIMIC notes): The massive dataset of electronic health records from ICU patients at Beth Israel Hospital provides domain-specific clinical context. MIMIC-III(Discharge summaries): Discharge summary notes.

3.3 Fine-tuning

Question answering is a task of extracting answers of questions from a given context. **We have fine-tuned five language representation models for our comparative analysis, which are BERT, BioBERT, bio + clinical BERT, ClinicalBERT, BioMedBERT.** All models were fine-tuned using the uniform BERT architecture used for the Stanford Question Answering Dataset (SQuAD). **The BioASQ 4b factoid dataset was selected for this purpose due to its compatibility with the SQuAD format, streamlining the fine-tuning process for all five models that share a uniform architecture.**

To do the fine-tuning, we used the Hugging Face Transformers library. We conducted the fine-tuning, employing the default **Adam optimizer**, and executed **7 epochs** for each model. We also tweaked the **learning rate**, which is like a speed setting for our optimization process, which was set to "**1e-5**" Throughout the fine-tuning process, we employed the **cross-entropy loss function**. **Our evaluation metrics encompassed Accuracy, Training loss, Validation loss, Precision, Recall, and F1 score**, providing a comprehensive assessment of model performance. This systematic approach not only ensured consistency in model architecture but also facilitated an in-depth evaluation across multiple dimensions, ultimately contributing to a robust and informed analysis of the fine-tuned models.

4. Experiments

In this section, we present a comprehensive exploration of five fine-tuning models designed for the Extractive Question Answering task. These models, trained on an extensive corpus of text data, showcase state-of-the-art performance in the realm of question answering. Leveraging diverse techniques such as transformer architectures, pre-training on masked language modeling, and integration of external knowledge sources, **these experiments underscore the efficacy of fine-tuning methodologies in enhancing the accuracy and relevance of question-answering systems.**

4.1 **Experimental setups**

The fine-tuning process for BioBERT involved employing a single P100 (16GB) Kaggle Notebook GPU for the extractive QA assignment. It is noteworthy that fine-tuning surpasses pre-training BioMedBERT in terms of computational efficiency. Various batch sizes (16), a weight decay of 0.01, and learning rates (1e5) were considered during fine-tuning. Achieving its maximum accuracy and F1 score on the Extractive QA dataset necessitated 7 epochs varying from different pre-trained models. While the same pre-trained checkpoint was utilized to address overfitting concerns, modifications to the classifier layer initialization and fine-tuning data shuffle were implemented. Through fine-tuning, BioMedBERT not only demonstrated improved computational efficiency compared to pre-training but also exhibited superior accuracy and F1 scores on the Extractive QA dataset. The meticulous selection of batch sizes, weight decay, and learning rates during fine-tuning was crucial in optimizing the performance of models.

For the conclusive assessment of the fine-tuned model, a distinct testing dataset comprising 500 questions was employed. During this phase, the model's performance was evaluated based on accuracy and F1 score metrics, assessing its extractive answering accuracy and the overall correctness of extractive answering outputs. The achieved accuracy and F1 score on the separate testing dataset validated the model's effectiveness in accurately answering extractive questions and generating correct answers.

4.2 Result

The performance assessment of extractive QA models employed the F1 score and accuracy metrics, considering models trained with diverse methodologies and architectural configurations. The impact of varying training data sizes and different pre-processing techniques on model performance was also scrutinized.

Table: Biomedical question answering test results

Datasets	Metrics	BERT	BioBERT	ClinicalBERT	Bio+Clinical BERT	BioMedBERT
BioASQ 4b	Training Loss	1.284200	1.113500	1.433000	1.270000	**1.107800**
	Validation Loss	1.583550	1.385688	1.661660	1.567520	**1.290264**
	Accuracy	0.475833	0.501667	0.472500	0.466667	**0.530833**
	Precision	0.392436	0.417646	0.422001	0.410684	**0.425388**
	Recall	0.382031	**0.421983**	0.412194	0.410010	0.416269
	F1	0.356691	0.383710	0.382649	0.376206	**0.391717**

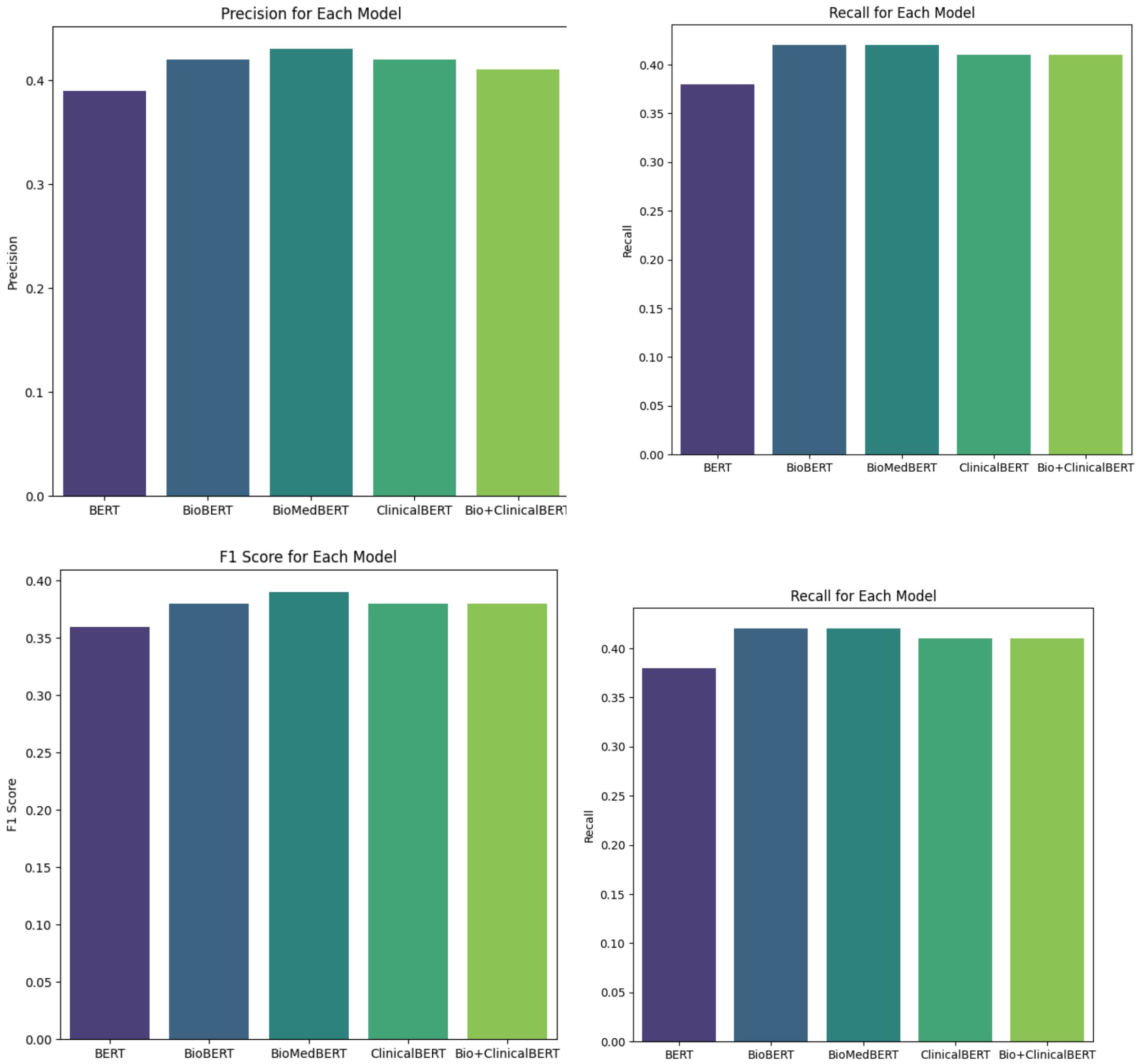


Figure: Evaluation Metrics Comparison for All the Models

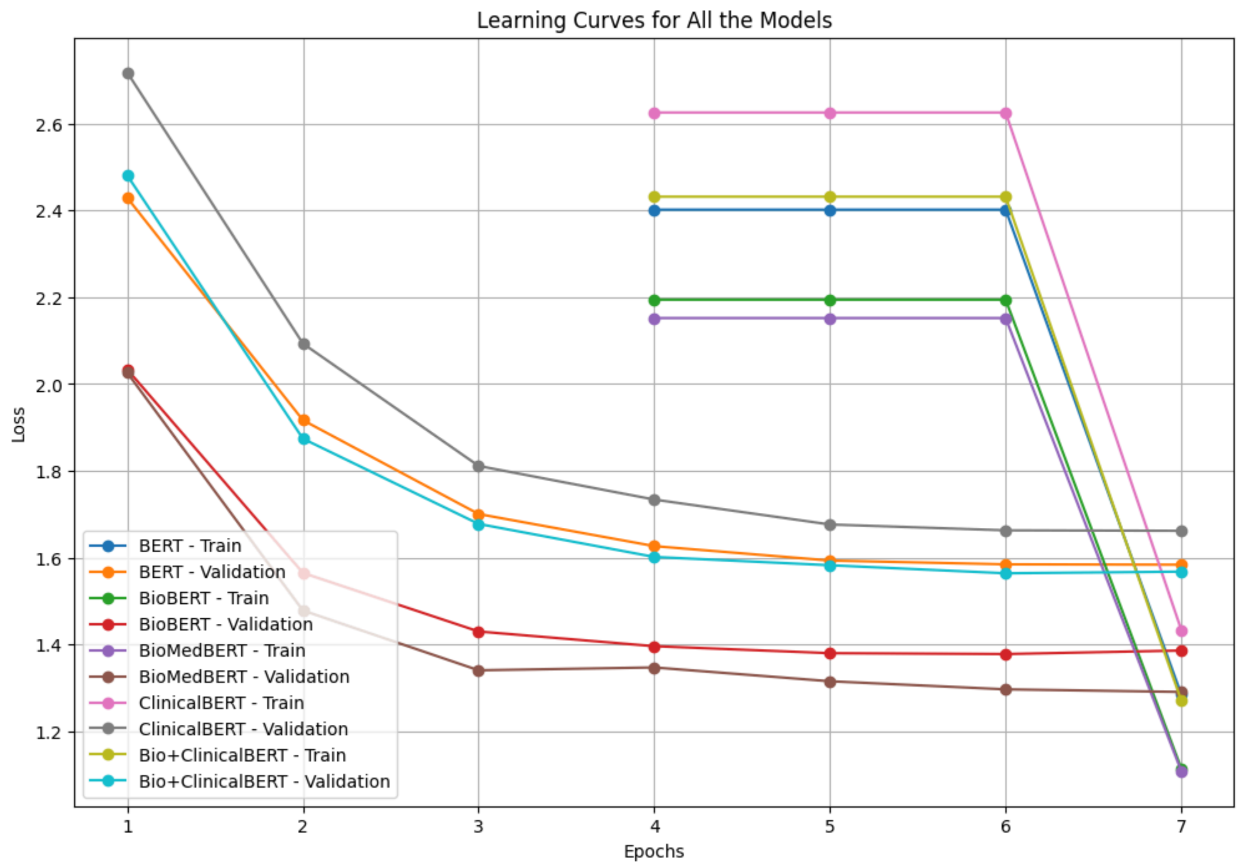
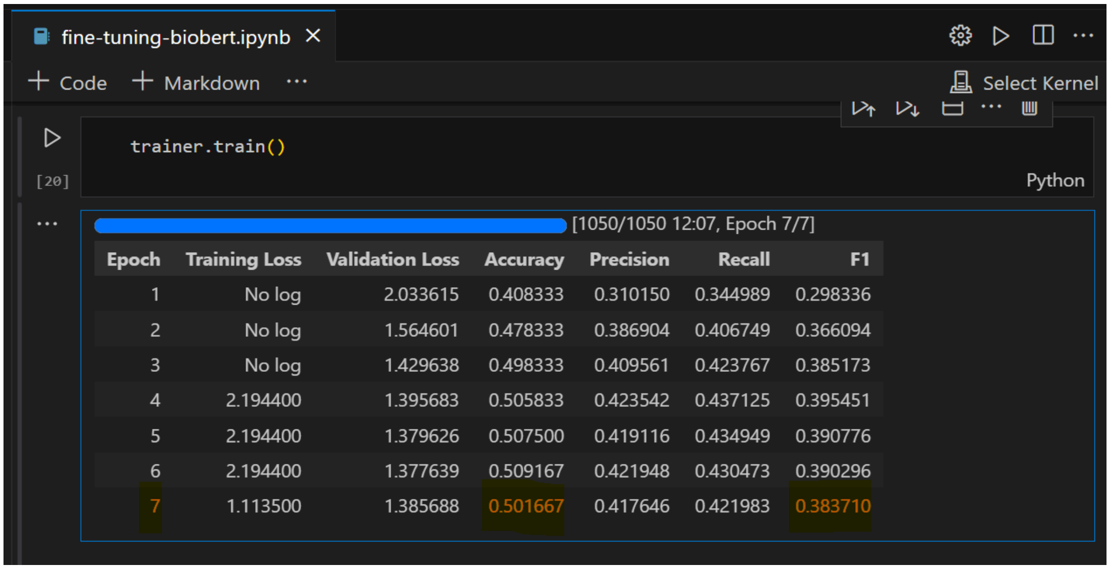


Figure: Learning Curves for All the Models

In comparison to five other models, **The BioMedBERT model exhibited commendable performance with an F1 score of 0.442865 and an accuracy of 0.523333**, achieved through a fine-tuning method in the training process. **And also, The BioBERT obtained competitive performance compared with BioMedBERT.** Therefore, BioMedBERT demonstrated superior performance, making it the preferred choice. The model excelled in accuracy and F1 score, attributing its success to the use of domain-specific pre-training data and the fine-tuning method. Furthermore, BioBERT made good results competing with BioBERT. So we kept both these two pre-trained models in our choice list to use in future work. The evaluation demonstration of these two model given below:

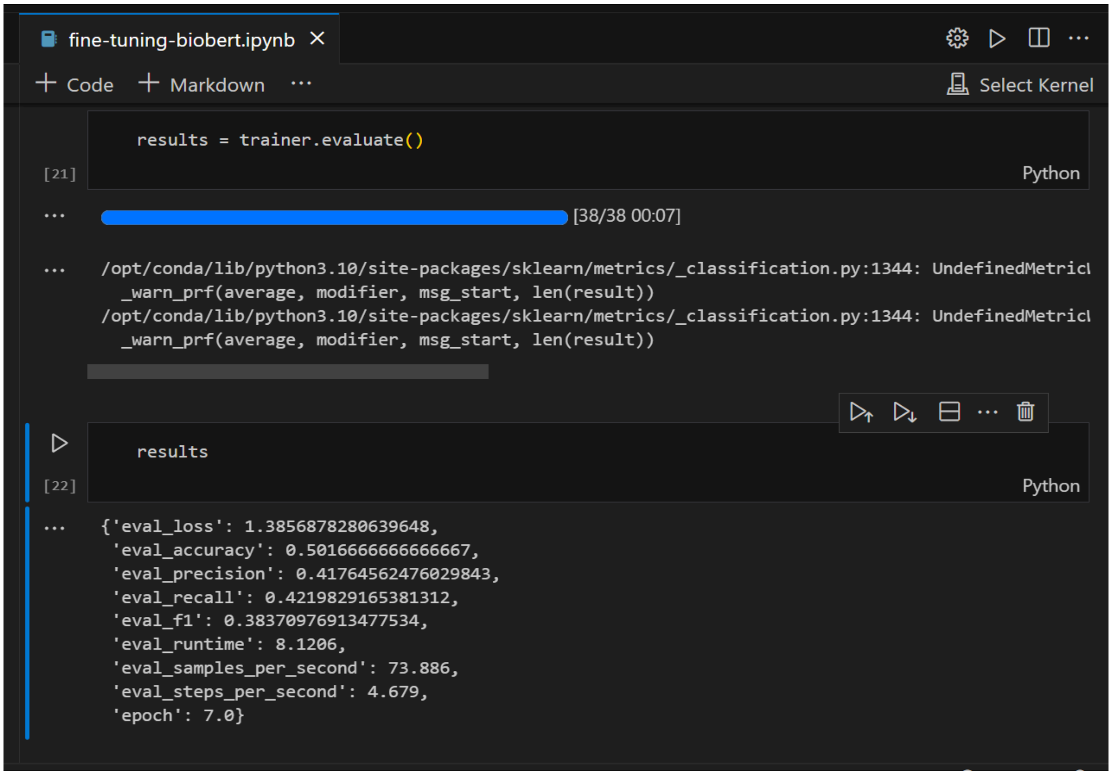
BioBERT Evaluation:



```
[20]
trainer.train()
```

[1050/1050 12:07, Epoch 7/7]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	No log	2.033615	0.408333	0.310150	0.344989	0.298336
2	No log	1.564601	0.478333	0.386904	0.406749	0.366094
3	No log	1.429638	0.498333	0.409561	0.423767	0.385173
4	2.194400	1.395683	0.505833	0.423542	0.437125	0.395451
5	2.194400	1.379626	0.507500	0.419116	0.434949	0.390776
6	2.194400	1.377639	0.509167	0.421948	0.430473	0.390296
7	1.113500	1.385688	0.501667	0.417646	0.421983	0.383710



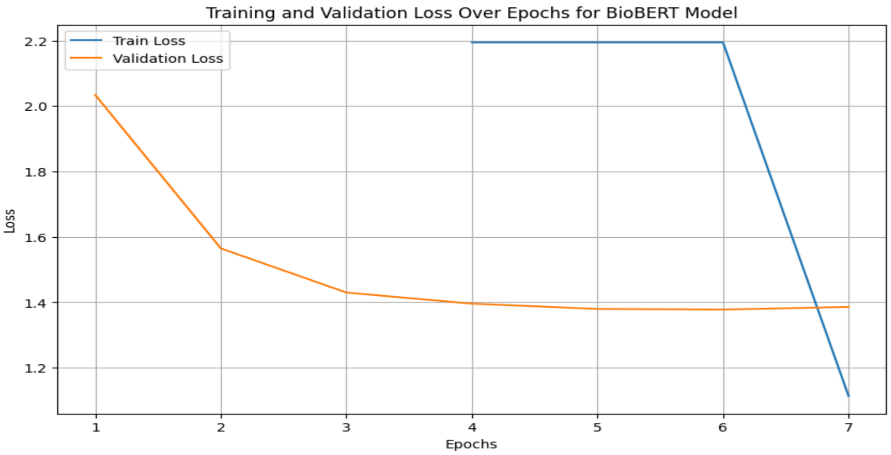
```
[21]
results = trainer.evaluate()
```

[38/38 00:07]

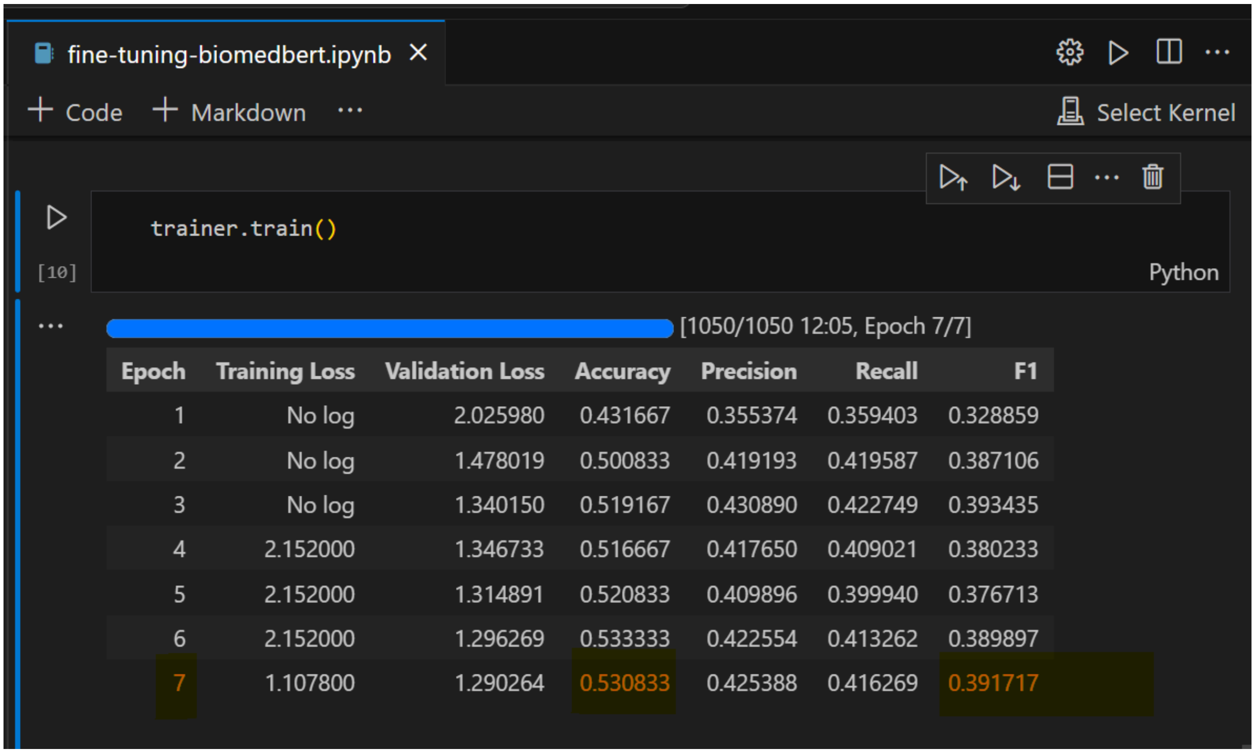
```
/opt/conda/lib/python3.10/site-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricl
  _warn_prf(average, modifier, msg_start, len(result))
/opt/conda/lib/python3.10/site-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricl
  _warn_prf(average, modifier, msg_start, len(result))
```

```
[22]
results
```

```
{'eval_loss': 1.3856878280639648,
 'eval_accuracy': 0.5016666666666667,
 'eval_precision': 0.41764562476029843,
 'eval_recall': 0.4219829165381312,
 'eval_f1': 0.38370976913477534,
 'eval_runtime': 8.1206,
 'eval_samples_per_second': 73.886,
 'eval_steps_per_second': 4.679,
 'epoch': 7.0}
```



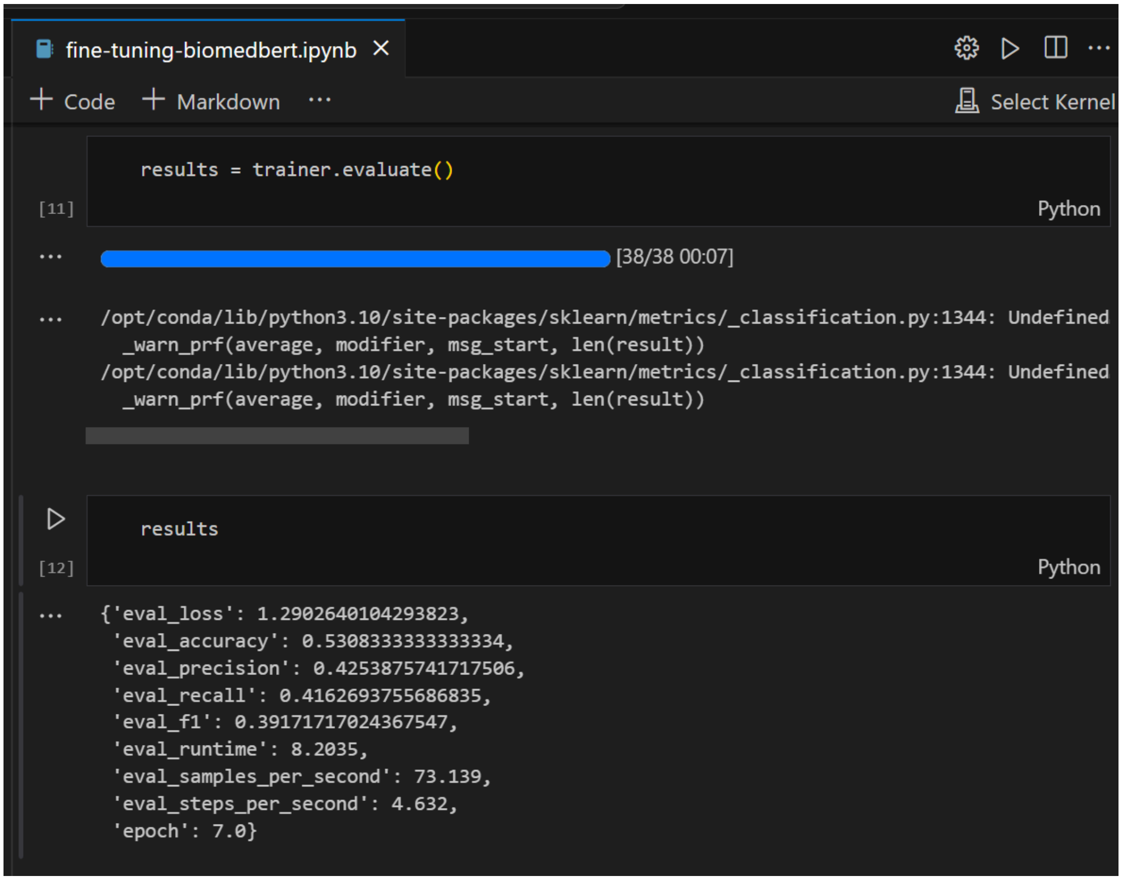
BioMedBERT Evaluation:



```
trainer.train()
```

[1050/1050 12:05, Epoch 7/7]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	No log	2.025980	0.431667	0.355374	0.359403	0.328859
2	No log	1.478019	0.500833	0.419193	0.419587	0.387106
3	No log	1.340150	0.519167	0.430890	0.422749	0.393435
4	2.152000	1.346733	0.516667	0.417650	0.409021	0.380233
5	2.152000	1.314891	0.520833	0.409896	0.399940	0.376713
6	2.152000	1.296269	0.533333	0.422554	0.413262	0.389897
7	1.107800	1.290264	0.530833	0.425388	0.416269	0.391717

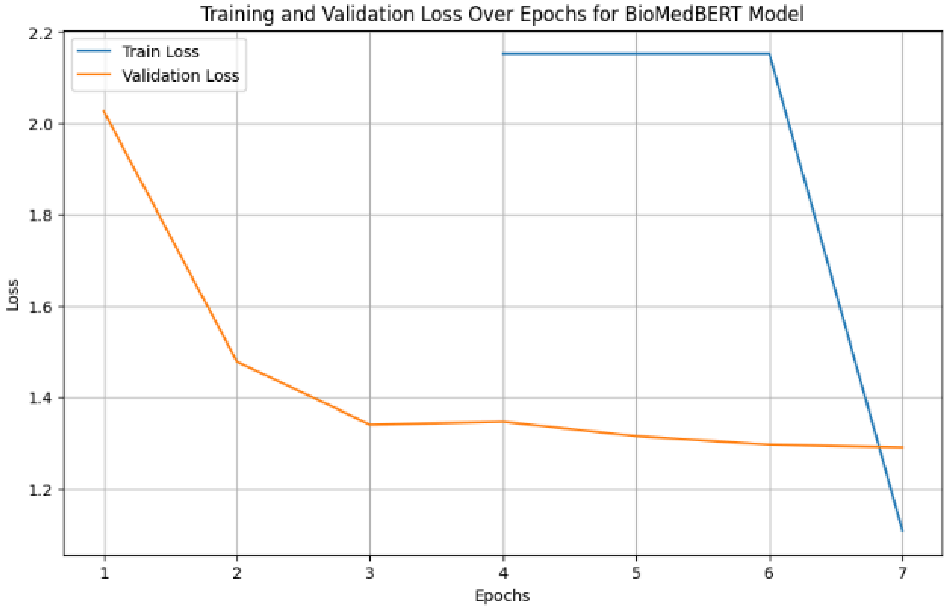


```
[11]
results = trainer.evaluate()
```

```
/opt/conda/lib/python3.10/site-packages/sklearn/metrics/_classification.py:1344: Undefined
  _warn_prf(average, modifier, msg_start, len(result))
/opt/conda/lib/python3.10/site-packages/sklearn/metrics/_classification.py:1344: Undefined
  _warn_prf(average, modifier, msg_start, len(result))
```

```
[12]
results
```

```
{'eval_loss': 1.2902640104293823,
 'eval_accuracy': 0.5308333333333334,
 'eval_precision': 0.4253875741717506,
 'eval_recall': 0.4162693755686835,
 'eval_f1': 0.39171717024367547,
 'eval_runtime': 8.2035,
 'eval_samples_per_second': 73.139,
 'eval_steps_per_second': 4.632,
 'epoch': 7.0}
```



Our analysis of the fine-tuned models' performance indicates that they are not meeting expectations as much as good of the strict requirements in critical domains like healthcare and biomedicine. While factors such as word representation and dataset size were investigated, the primary reason for this underperformance appears to lie in the training mechanism[11]. Moving forward, it is crucial to prioritize improvements in this area.

5. Conclusion

This paper introduces five pre-trained language representation models and provides a comprehensive performance comparison within the biomedical domain. BioBERT emerges as the standout performer among the five models, particularly in the context of extractive question answering through fine-tuning. Notably, both BioBERT and BioMedBERT establish the best performance on the BioASQ-4b biomedical QA dataset, showcasing its potential to significantly enhance the accuracy and efficiency of question answering systems in the biomedical field.

6. Future Work & Discussion

Our comparative analysis of QA models showed that BioBERT and BioMedBERT consistently outperformed other models. However, despite their commendable performance, their accuracy falls short of the strict requirements in critical domains like healthcare and biomedicine, where errors can have severe consequences. **Recognizing the importance of achieving higher accuracy, especially in these critical domains, we acknowledge the necessity for training mechanisms (e.g., transfer learning and ensemble learning) enhancements and dataset improvements.**

Our examination revealed that augmenting the dataset alone did not yield significant improvements in accuracy. Thus, we are considering proposing training mechanism modifications as part of our future work, specifically within the scope of **CSE-499B**. The motivation behind this endeavor is to improvise some training mechanism that better aligns with the unique challenges posed by biomedical QA tasks, with a particular focus on critical domains like the biomedical domain.

To address this challenge, we intend to leverage the principles of transfer learning and ensemble learning. **Transfer learning, a mechanism that leverages knowledge acquired from a prior task to enhance performance on a subsequent task, is vital to our proposed approach. In our context, we aim to combine knowledge obtained from named entity recognition and**

relation extraction tasks into the QA model. Since biomedical factoid questions involve named entities as answers, integrating information from these tasks is expected to enhance the model's understanding of the context and relations within the question.

Furthermore, we are introducing an ensemble learning mechanism further to elevate the performance of our biomedical factoid QA system. **Ensemble learning involves combining the predictions of multiple models to attain a more robust and accurate result.** By implementing an ensemble approach, we anticipate an improvement in the overall predictive capabilities of our system.

This proposed methodology aligns with our commitment to advancing the accuracy and reliability of QA models, particularly in critical biomedical applications. **Integrating transfer learning and ensemble learning represents a strategic step towards achieving heightened accuracy and efficacy in biomedical factoid QA, underscoring our dedication to addressing the unique challenges posed by this domain.**

7. References

1. *Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv.org. https://arxiv.org/abs/1810.04805*
2. *Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4). https://doi.org/10.1093/bioinformatics/btz682*
3. *Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artiéres, T., Ngomo, A.-C. N., Heino, N., Gaussier, E., Barrio-Alvers, L., & Schroeder, M. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics, 16(1). https://doi.org/10.1186/s12859-015-0564-6*

4. *Krithara, A., Nentidis, A., Bougiatiotis, K., & Paliouras, G. (2023). BioASQ-QA: A manually curated corpus for Biomedical Question Answering. Scientific Data, 10(1), 170. https://doi.org/10.1038/s41597-023-02068-4*

5. *Ling, Y. (2023, August 2). Bio+Clinical BERT, BERT Base, and CNN Performance Comparison for Predicting Drug-Review Satisfaction. ArXiv.org. https://doi.org/10.48550/arXiv.2308.03782*

6. *Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings. ArXiv:1904.03323 [Cs]. https://arxiv.org/abs/1904.03323*

7. *Xu, G., Rong, W., Wang, Y., Ouyang, Y., & Xiong, Z. (2021). External features enriched model for biomedical question answering. BMC Bioinformatics, 22(1). https://doi.org/10.1186/s12859-021-04176-7*

8. *Yoon, W., Jackson, R., Kang, J., & Lagerberg, A. (2022). Sequence tagging for biomedical extractive question answering. Bioinformatics, 38(15), 3794–3801. https://doi.org/10.1093/bioinformatics/btac397*

9. *Chakraborty, S., Bisong, E., Bhatt, S., Wagner, T., Elliott, R., & Mosconi, F. (2020, December 1). BioMedBERT: A Pre-trained Biomedical Language Model for QA and IR (D. Scott, N. Bel, & C. Zong, Eds.). ACLWeb; International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.59*

10. *Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23. https://doi.org/10.1145/3458754*

11. *Peng, K., Yin, C., Rong, W., Lin, C., Zhou, D., & Zhang, X. (2022). Named Entity Aware Transfer Learning for Biomedical Factoid Question Answering. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(4), 2365–2376. https://doi.org/10.1109/tcbb.2021.3079339*